R E P U B L I Q U E F R A N Ç A I S

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION

CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 18 JUIL 2003

Pour le Directeur général de l'Institut
national de la propriété industrielle
Le Chef du Département des brevets

Martine PLANCHE

DOCUMENT DE PRIORITÉ

PRÉSENTÉ OU TRANSMIS CONFORMÉMENT À LA RÈGLE 17.1.a) OU b)

INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
SIEGE
26 bis, rue de Saint Petersbourg
75800 PARIS cedex 08
Téléphone : 33 (0)1 53 04 53 04
Télécopie : 33 (0)1 53 04 45 23
www.inpi.fr

1er dépôt

INPI
INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 33 (1) 53 04 53 04 Télécopie : 33 (1) 42 94 86 54

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11354*02

REQUÊTE EN DÉLIVRANCE
page 1/2

BR1

Cet imprimé est à remplir lisiblement à l'encre noire DB 540 @ W / 010801

Réservé à l'INPI

REMISE DES PIÈCES DATE 8 JUIL 2002
LIEU 69 INPI LYON
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0208548
DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI 08 JUIL. 2002

Vos références pour ce dossier *(facultatif)* 70416BFR23 JMT/VF

1 NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE

Cabinet BEAU DE LOMENIE
51, Avenue Jean Jaurès
B.P. 7073

69301 LYON CEDEX 07

Confirmation d'un dépôt par télécopie ☐ N° attribué par l'INPI à la télécopie

2 NATURE DE LA DEMANDE — Cochez l'une des 4 cases suivantes

Demande de brevet	☒	
Demande de certificat d'utilité	☐	
Demande divisionnaire	☐	
Demande de brevet initiale	N°	Date
ou demande de certificat d'utilité initiale	N°	Date
Transformation d'une demande de brevet européen *Demande de brevet initiale*	☐ N°	Date

3 TITRE DE L'INVENTION (200 caractères ou espaces maximum)

PROCEDE ET APPAREIL POUR AFFECTER UNE CLASSE SONORE A UN SIGNAL SONORE

4 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE FRANÇAISE

Pays ou organisation
Date N°
Pays ou organisation
Date N°
Pays ou organisation
Date N°

☐ **S'il y a d'autres priorités, cochez la case et utilisez l'imprimé «Suite»**

5 DEMANDEUR (Cochez l'une des 2 cases) ☒ **Personne morale** ☐ **Personne physique**

Nom ou dénomination sociale		ECOLE CENTRALE DE LYON
Prénoms		
Forme juridique		Etablissement Public à Caractère Scientifique, Culurel et Professionnel
N° SIREN		
Code APE-NAF		
Domicile ou siège	Rue	36, Avenue Guy de Collongue B.P. 163
	Code postal et ville	6 9 1 3 1 ECULLY CEDEX
	Pays	FRANCE
Nationalité		française
N° de téléphone *(facultatif)*		N° de télécopie *(facultatif)*
Adresse électronique *(facultatif)*		

☐ **S'il y a plus d'un demandeur, cochez la case et utilisez l'imprimé «Suite»**

Remplir impérativement la 2ème page

1er dépôt

INPI
INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

REQUÊTE EN DÉLIVRANCE
page 2/2

BR2

Réservé à l'INPI

REMISE DES PIÈCES
DATE 8 JUIL 2002
LIEU 69 INPI LYON
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0208548

DB 540 @ W / 010801

Vos références pour ce dossier : *(facultatif)*		70416BFR23 JMT/VF
6 **MANDATAIRE** *(s'il y a lieu)*		
Nom		THIBAULT
Prénom		Jean-Marc
Cabinet ou Société		Cabinet BEAU DE LOMENIE
N °de pouvoir permanent et/ou de lien contractuel		
Adresse	Rue	51, Avenue Jean Jaurès
	Code postal et ville	6 9 3 0 1 LYON CEDEX 07
	Pays	FRANCE
N° de téléphone *(facultatif)*		04 72 76 85 30
N° de télécopie *(facultatif)*		04 78 69 86 82
Adresse électronique *(facultatif)*		
7 **INVENTEUR (S)**		**Les inventeurs sont nécessairement des personnes physiques**
Les demandeurs et les inventeurs sont les mêmes personnes		☐ Oui ☒ Non : **Dans ce cas remplir le formulaire de Désignation d'inventeur(s)**
8 **RAPPORT DE RECHERCHE**		**Uniquement pour une demande de brevet (y compris division et transformation)**
Établissement immédiat ou établissement différé		☒ ☐
Paiement échelonné de la redevance *(en deux versements)*		**Uniquement pour les personnes physiques effectuant elles-mêmes leur propre dépôt** ☐ Oui ☐ Non
9 **RÉDUCTION DU TAUX DES REDEVANCES**		**Uniquement pour les personnes physiques** ☐ Requise pour la première fois pour cette invention *(joindre un avis de non imposition)* ☐ Obtenue antérieurement à ce dépôt pour cette invention *(joindre une copie de la décision d'admission à l'assistance gratuite ou indiquer sa référence)* : **AG**
Si vous avez utilisé l'imprimé «Suite», indiquez le nombre de pages jointes		

10 **SIGNATURE DU DEMANDEUR OU DU MANDATAIRE (Nom et qualité du signataire)** Jean-Marc THIBAULT Conseil en P.I. - N° 94-0312	**VISA DE LA PRÉFECTURE OU DE L'INPI** M. DIEZ

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire. Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.

La présente invention concerne le domaine de la classification d'un signal sonore en des classes acoustiques reflétant une sémantique.

L'objet de l'invention concerne plus précisément le domaine de l'extraction automatique d'un signal sonore, d'informations sémantiques tels que musique, parole, bruit, silence, homme, femme, musique rock, jazz, etc.

Dans l'état de la technique, la profusion de documents multimédias requiert une indexation nécessitant une intervention humaine importante, ce qui constitue une opération coûteuse et longue à mener à bien. Par conséquent, l'extraction automatique d'informations sémantiques constitue une aide précieuse permettant de faciliter et d'accélérer le travail de l'analyse et de l'indexation.

Dans de nombreuses applications, la segmentation et la classification sémantique d'une bande sonore constituent fréquemment des opérations nécessaires avant d'envisager d'autres analyses et traitements sur le signal sonore.

Une application connue nécessitant la segmentation et la classification sémantique concerne les systèmes de reconnaissance automatique de la parole appelés aussi systèmes de dictée vocale adaptés pour transcrire en texte une bande de paroles. Une segmentation et une classification de la bande sonore en des segments musique/parole sont des étapes indispensables pour un niveau de performances acceptables.

L'utilisation d'un système de reconnaissance automatique de la parole pour une indexation par le contenu de documents audiovisuels comme par exemple les journaux télévisés, nécessite d'éliminer les segments de non parole pour diminuer le taux d'erreur. De plus, si une connaissance a priori du genre du locuteur (homme ou femme) est disponible, l'utilisation d'un système de reconnaissance automatique de la parole permet d'aboutir à une amélioration importante des performances.

Une autre application connue ayant recours à la segmentation et à la classification sémantique d'une bande sonore concerne des systèmes de statistiques et de surveillance. En effet, pour des questions du respect du droit d'auteur ou du respect du quota du temps de parole, des organismes de régulation et de contrôle comme le CSA ou la SACEM en France, doivent s'appuyer sur des comptes rendus précis, par exemple sur la durée du temps de parole par homme politique dans les chaînes de télévision pour le CSA et le titre et la durée des chansons émises par les

radios pour la SACEM. La mise en place d'un système automatique de statistiques et de surveillance s'appuie au préalable sur une segmentation et une classification d'une bande sonore musique/parole.

Une autre application possible a trait au système de résumé ou de filtrage automatique de programmes audiovisuels. Pour de nombreuses applications, comme par exemple la téléphonie mobile ou la vente de programmes audiovisuels par correspondance, il apparaît nécessaire de résumer éventuellement selon le centre d'intérêt d'un utilisateur, un programme audiovisuel de deux heures en une compilation de moments forts de quelques minutes. Un tel résumé peut être réalisé soit off-line, c'est-à-dire qu'il s'agit d'un résumé préalablement calculé qui est associé au programme d'origine, soit on-line, c'est-à-dire qu'il s'agit d'un filtrage du programme audiovisuel permettant de conserver uniquement les moments forts d'un programme en mode de diffusion ou streaming. Les moments forts sont fonction du programme audiovisuel et du centre d'intérêt d'un utilisateur. Par exemple, dans un match de football, un moment fort est celui où il y a une action de but. Pour un film d'action, un moment fort correspond à des combats, à des poursuites, etc. Ces moments forts se traduisent le plus souvent en des percussions sur la bande sonore. Pour les identifier, il est intéressant de s'appuyer sur une segmentation et une classification de la bande sonore en des segments ayant une certaine propriété ou non.

En l'état de la technique, il existe divers systèmes de classification d'un signal sonore. Par exemple, le document WO 98 27 543 décrit une technique de classification d'un signal sonore en musique ou parole. Ce document prévoit d'étudier les différents paramètres mesurables du signal sonore tel que l'énergie de modulation à 4Hz, le flux spectral, la variation du flux spectral, le taux de passage par zéro, etc. Ces paramètres sont extraits pour une fenêtre d'une seconde ou une autre durée, pour définir la variation du flux spectral ou une trame comme le taux de passage par zéro. Ensuite, en utilisant différents classificateurs, comme par exemple le classificateur basé sur le mélange des lois Gaussiennes ou un classificateur du Plus Proche Voisin, un taux d'erreur de l'ordre de 6 % est obtenu. L'apprentissage des classificateurs a été réalisé sur trente six minutes et le test sur quatre minutes. Ces résultats montrent que la technique proposée nécessitent une base d'apprentissage

d'une taille importante pour aboutir à un taux de reconnaissance de 95 %. Si cela est possible avec quarante minutes de documents audiovisuels, cette technique apparaît difficilement envisageable pour des applications où les données à classifier ont une taille importante avec un niveau haut de variabilité résultant des différentes sources des documents avec des niveaux de bruits et de résolution différents pour chacune de ces sources.

Le brevet US 5 712 953 décrit un système utilisant la variation par rapport au temps du premier moment du spectre relatif à la fréquence pour la détection du signal de musique. Ce document suppose que cette variation est très faible pour la musique contrairement à d'autres signaux non musicaux. Malheureusement, les différents types de musique n'ont pas la même structuration de sorte qu'un tel système présente des performances insuffisantes comme par exemple pour le RAP.

La demande de brevet européen 1 100 073 propose une classification du signal sonore en différentes catégories en utilisant dix-huit paramètres comme par exemple la moyenne et la variance de la puissance du signal, la puissance des moyennes fréquences, etc. Une quantification vectorielle est réalisée et la distance de Mahalanobis est utilisée pour la classification. Il apparaît que l'utilisation de la puissance du signal n'est pas stable car les signaux provenant de différentes sources sont toujours enregistrés avec différents niveaux de puissance spectrale. Par ailleurs, l'utilisation des paramètres, comme la puissance de basses fréquences ou hautes fréquences, pour la discrimination entre la musique et la parole est une limitation sérieuse compte tenu de l'extrême variation, à la fois de la musique et de la parole. Enfin, le choix d'une distance appropriée pour des vecteurs de dix-huit paramètres non homogènes n'est pas évident car il s'agit d'affecter des poids différents à ces paramètres en fonction de leur importance.

L'objet de l'invention vise donc à remédier aux inconvénients énoncés ci-dessus en proposant une technique permettant de réaliser une classification du signal sonore en des classes sémantiques avec un taux de reconnaissance élevé tout en nécessitant une durée réduite d'apprentissage.

Pour atteindre un tel objectif, le procédé selon l'invention concerne un procédé pour affecter au moins une classe sonore à un signal sonore, comprenant les étapes suivantes :

- diviser le signal sonore en des segments temporels présentant une durée déterminée,
- extraire les paramètres fréquentiels du signal sonore dans chacun des segments temporels,
- regrouper les paramètres fréquentiels dans des fenêtres temporelles présentant une durée déterminée supérieure à la durée des segments temporels,
- extraire de chaque fenêtre temporelle, des composantes caractéristiques,
- et en considération des composantes caractéristiques extraites et à l'aide d'un classificateur, identifier la classe sonore de chaque fenêtre temporelle du signal sonore.

Un autre objet de l'invention est de proposer un appareil pour affecter au moins une classe sonore à un signal sonore comprenant :

- des moyens pour diviser le signal sonore en des segments temporels présentant une durée déterminée,
- des moyens pour extraire les paramètres fréquentiels du signal sonore dans chacun des segments temporels,
- des moyens pour regrouper les paramètres fréquentiels dans des fenêtres temporelles présentant une durée déterminée supérieure à la durée des segments temporels,
- des moyens pour extraire de chaque fenêtre temporelle, des composantes caractéristiques,
- et des moyens pour identifier la classe sonore des fenêtres temporelles du signal sonore en considération des composantes caractéristiques extraites et à l'aide d'un classificateur.

Diverses autres caractéristiques ressortent de la description faite ci-dessous en référence aux dessins annexés qui montrent, à titre d'exemples non limitatifs, des formes de réalisation de l'objet de l'invention.

La **Fig. 1** est un schéma synoptique montrant un appareil de mise en œuvre du procédé de classification d'un signal sonore conforme à l'invention.

La **Fig. 2** est un schéma illustrant une étape caractéristique du procédé selon l'invention, à savoir de transformation.

La **Fig. 3** est un schéma illustrant une autre étape caractéristique de l'invention.

La **Fig. 4** illustre une étape de classification du signal sonore selon l'invention.

La **Fig. 5** est un schéma illustrant un exemple de réseau de neurones utilisé dans le cadre de l'invention.

Tel que cela apparaît plus précisément à la **Fig. 1**, l'objet de l'invention concerne un appareil **1** permettant de classifier un signal sonore **S** de tous types en des classes sonores. En d'autres termes, le signal sonore **S** est découpé en des segments qui sont étiquetés en fonction de leur contenu. Les étiquettes associées à chaque segment comme par exemple musique, parole, bruit, homme, femme, etc. réalisent une classification du signal sonore en des catégories sémantiques ou classes sonores sémantiques.

Conformément à l'invention, le signal sonore **S** à classifier est appliqué à l'entrée de moyens de segmentation **10** permettant de diviser le signal sonore **S** en des segments temporels **T** présentant chacun une durée déterminée. De préférence, les segments temporels **T** présentent tous une même durée comprise de préférence entre dix et trente ms. Dans la mesure où chaque segment temporel **T** présente une durée de quelques millisecondes, il peut être considéré que le signal est stationnaire, de sorte qu'il peut être appliqué par la suite, des transformations qui changent le signal temporel dans le domaine fréquentiel. Différents types de segments temporels peuvent être utilisés comme par exemple des fenêtres rectangulaires simples, fenêtres de Hanning ou de Hamming.

L'appareil **1** comporte ainsi des moyens d'extraction **20** permettant d'extraire les paramètres fréquentiels du signal sonore dans chacun des segments temporels **T**. L'appareil **1** comporte également des moyens **30** pour regrouper ces paramètres fréquentiels dans des fenêtres temporelles **F** présentant une durée déterminée supérieure à la durée des segments temporels **T**.

Selon une caractéristique préférée de réalisation, les paramètres fréquentiels sont regroupés dans des fenêtres temporelles **F** de durée supérieure à 0,3 seconde et de préférence comprise entre 0, 5 et 2 secondes. Le choix de la taille de la fenêtre temporelle **F** est déterminé pour pouvoir discriminer deux fenêtres différentes acoustiquement comme par exemple parole, musique, homme, femme, silence, etc. Si la fenêtre temporelle **F** est courte de quelques dizaines de millisecondes par

exemple, des changements acoustiques locaux de type changement de volume, changement d'instrument de musique, début ou fin d'un mot peuvent être détectés. Si la fenêtre est large, par exemple de quelques centièmes de millisecondes par exemple, les changements détectables seront des changements plus généraux du type changement de rythme de musique ou rythme de parole par exemple.

L'appareil **1** comporte également des moyens d'extraction **40** permettant d'extraire de chaque fenêtre temporelle **F** des composantes caractéristiques. En considération de ces composantes caractéristiques extraites et à l'aide d'un classificateur **50**, des moyens d'identification **60** permettent d'identifier la classe sonore de chaque fenêtre temporelle **F** du signal sonore **S**.

La description qui suit décrit une variante préférée de réalisation d'une méthode de classification d'un signal sonore.

Selon une caractéristique préférée de réalisation, pour passer du domaine temporel au domaine fréquentiel, les moyens d'extraction **20** utilisent la Transformée de Fourier Discrète dans le cas d'un signal sonore échantillonné, notée par la suite TFD. La Transformée de Fourier Discrète donne pour une série temporelle de valeurs d'amplitude du signal, une série de valeurs de spectres de fréquence. L'équation de la Transformée de Fourier Discrète est la suivante :

$$X_N(n) = \sum_{k=0}^{N-1} x(k) e^{-j2\pi kn/N}$$

où x(k) est le signal dans le domaine temporel.

Le terme |X(n)| est appelé *spectre d'amplitude*, il exprime la répartition fréquentielle de l'amplitude du signal x(k).

Le terme arg[X(n)] est appelé *spectre de phase*, il exprime la répartition fréquentielle de la phase du signal x(k).

Le terme $|X(n)|^2$ est appelé *spectre d'énergie*, exprimant la répartition fréquentielle de l'énergie du signal x(k).

Les valeurs largement utilisées sont les valeurs de spectre d'énergie.

En conséquence, pour une série de valeurs temporelles de l'amplitude du signal x(k) d'un segment temporel **T**, il est obtenu une série X_i des valeurs du spectre de

fréquence dans une plage de fréquences comprise entre une fréquence minimale et une fréquence maximale. La collection de ces valeurs ou paramètres fréquentiels est appelée « vecteur de TFD » ou vecteur spectral. Chaque vecteur $\mathbf{X}_i$ correspond au vecteur spectral pour chaque segment temporel **T**, avec **i** allant de **1** à **n**.

Selon une caractéristique préférée de réalisation, une opération de transformation ou de filtrage est effectuée sur les paramètres fréquentiels préalablement obtenus par l'intermédiaire de moyens de transformation **25** interposés entre les moyens d'extraction **20** et les moyens de regroupement **30**. Tel que cela apparaît plus précisément sur la **Fig. 2**, cette opération de transformation permet à partir du vecteur spectral $\mathbf{X}_i$, de générer un vecteur de caractéristiques transformées $\mathbf{Y}_i$. La transformation est donnée par la formule y_i avec les variables, limite1, limite2 et aj qui définissent précisément la transformation.

La transformation peut être du type identité de sorte que le vecteur de caractéristiques $\mathbf{X}_i$ ne change pas. Selon cette transformation, limite1 et limite2 sont égaux à j et le paramètre aj est égal à 1. Le vecteur spectral $\mathbf{X}_i$ est égal $\mathbf{Y}_i$.

La transformation peut être une transformation moyenne de deux fréquences adjacentes. Selon ce type de transformation, il peut être obtenu la moyenne de deux spectres de fréquences adjacentes. Par exemple, il peut être choisi limite1 est égal à j et limite2 est égal à j+1 et aj est égal à 0,5.

La transformation utilisée peut être une transformation suivant une approximation de l'échelle de Mel. Cette transformation peut être obtenue en faisant varier les variables limite1 et limite2 sur les valeurs suivantes :

0, 1, 2, 3, 4, 5, 6, 8, 9,10, 12, 15, 17, 20, 23, 27,31, 37, 40, avec

$$a_j = \frac{1}{|\mathrm{limite1} - \mathrm{limite2}|}$$

Par exemple, en choisissant *limite1* et *limite2* comme indiqué ci-dessous il peut être obtenu un vecteur **Y** de dimension 20, à partir d'un vecteur brut **X** de dimension 40, en utilisant l'équation décrite dans la **Fig. 2**.

limite1=0 → limite2=1

limite1=1 → limite2=2

limite1=2 → limite2=3

limite1=3 → limite2=4

limite1=4 → limite2=5

limite1=5 → limite2=6

limite1=6 → limite2=8

limite1=8 → limite2=9

limite1=9 → limite2=10

limite1=10 → limite2=12

limite1=12 → limite2=15

limite1=15 → limite2=17

limite1=17 → limite2=20

limite1=20 → limite2=23

limite1=23 → limite2=27

limite1=27 → limite2=31

limite1=31 → limite2=37

limite1=37 → limite2=40

Les transformations sur le vecteur spectral $\mathbf{X}_i$ sont plus ou moins importantes selon l'application, c'est-à-dire en fonction des classes sonores à classifier. Des exemples de choix de cette transformation seront données dans la suite de la description.

Tel que cela ressort de la description qui précède, le procédé selon l'invention consiste à extraire de chaque fenêtre temporelle **F**, des composantes caractéristiques permettant d'obtenir une description du signal sonore sur cette fenêtre présentant une durée relativement large. Ainsi, pour les vecteurs $\mathbf{Y}_i$ de chaque fenêtre temporelle **F**, les composantes caractéristiques calculées peuvent être la moyenne, la variance, le moment, le paramètre du suivi des fréquences ou le taux de passage par silence. L'estimation de ces composantes caractéristiques est effectuée selon la formule suivante :

$$\vec{w}_i = \begin{pmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{iN} \end{pmatrix} \vec{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{iN} \end{pmatrix} \vec{v}_i = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iN} \end{pmatrix} \vec{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN} \end{pmatrix}$$

où $\vec{\mu}_i$ est le vecteur moyen, $\vec{v}_i$ le vecteur de variance, $\vec{x}_i$ étant le vecteur de caractéristiques qui n'est autre que le vecteur spectral filtré décrit précédemment pour constituer des fenêtres temporelles **F**.

$\mu_{ij} = \frac{1}{M_i} \sum_{l=1}^{M_i} x_{lj}$ $\quad j = 1, \cdots, N$ où j correspond à la bande de fréquence dans le vecteur spectral $\vec{x}$, l correspond au temps, ou l'instant pour lequel le vecteur est extrait (segment temporel T), N est le nombre d'éléments dans le vecteur (ou le nombre de bande de fréquence), M_i correspond au nombre de vecteur à étudier leurs statistiques (fenêtre temporelle F), i dans μ_{ij} correspond à l'instant de la fenêtre temporelle F pour laquelle μ_{ij} est calculée, j correspond à la bande de fréquence.

$v_{ij} = \frac{1}{M_i} \sum_{l=1}^{M_i} (x_{lj} - \mu_{ij})^2$ $\quad j = 1, \cdots, N$

où j correspond à la bande de fréquence dans le vecteur spectral $\vec{x}$ et dans le vecteur moyen $\vec{\mu}$, l correspond au temps, ou l'instant pour lequel le vecteur $\vec{x}$ est extrait (segment temporel T), N est le nombre d'éléments dans le vecteur (ou le nombre de bande de fréquence), M_i correspond au nombre de vecteur à étudier leurs statistiques (fenêtre temporelle F), i dans μ_{ij} et v_{ij} correspond à l'instant de la fenêtre temporelle F pour laquelle $\vec{\mu}$ et $\vec{v}$ sont calculées, j correspond à la bande de fréquence.

Le moment qui peut être important pour la description du comportement des données est calculé de la manière suivante :

$w_{ij} = \frac{1}{M_i} \sum_{l=1}^{M_i} (x_{lj} - \mu_{ij})^n$ $\quad j = 1, \cdots, N$, les indices i, j, N, l, M_i sont expliqués pour la variance, et $n > 2$.

Le procédé selon l'invention permet également de déterminer comme composantes caractéristiques, le paramètre **SF** permettant de suivre les fréquences. En effet, il a été constaté que pour la musique, il existait une certaine continuité de fréquences, c'est-à-dire que les fréquences les plus importantes dans le signal, c'est-à-dire celles qui concentrent le plus d'énergie restent les mêmes pendant un certain temps, tandis que pour la parole ou pour le bruit (non harmonique) le changement des fréquences les plus importantes se fait d'une manière plus rapide. A partir de ce constat, il est proposé de faire un suivi de plusieurs fréquences en même temps selon un intervalle de précision par exemple 200 Hz. Ce choix est motivé par le fait que les fréquences les plus importantes dans une musique changent mais d'une manière graduelle. L'extraction de ce paramètre de suivi de fréquences SF se fait de la manière suivante. Pour chaque vecteur Y_i de Transformée de Fourier Discrète, il est procédé à l'identification par exemple des cinq fréquences les plus importantes. Si l'une de ces fréquences ne figure plus dans les cinq fréquences les plus importantes du vecteur de Transformée de Fourier Discrète, dans une bande de 100 Hz, une coupure est signalée. Le nombre de coupures dans chaque fenêtre temporelle **F** est compté, ce qui définit le paramètre de suivi de fréquences **SF**. Ce paramètre **SF** pour les segments de musique est clairement inférieur à celui de la parole ou du bruit. Aussi, un tel paramètre est intéressant pour une discrimination entre la musique et la parole.

Selon une autre caractéristique de l'invention, le procédé consiste à définir comme composante caractéristique, le taux de passage par silence **TPPS**. Ce paramètre consiste à compter dans une fenêtre de taille fixée, par exemple de deux secondes, le nombre de fois où l'énergie arrive au seuil de silence. En effet, il doit être considéré que l'énergie du signal sonore pendant l'élocution d'un mot est normalement élevé alors qu'elle diminue sous le seuil de silence entre les mots. L'extraction du paramètre est effectué de la manière suivante. Pour chaque 10 ms du signal, l'énergie du signal est calculée. La dérivée de l'énergie est calculée par rapport au temps, soit l'énergie de **T+1** moins l'énergie à l'instant **T**. Puis dans une fenêtre de 2 secondes, le nombre de fois où la dérivée de l'énergie dépasse un certain seuil est comptée.

Tel que cela apparaît plus précisément à la **Fig. 3**, les paramètres extraits de chaque fenêtre temporelle **F** définissent un vecteur de caractéristiques $\mathbb{Z}$. Ce vecteur de caractéristiques $\mathbb{Z}$ est donc la concaténation des composantes caractéristiques définies à savoir les vecteurs moyens, variances et moments, ainsi que le suivi des fréquences **SF** et le taux de passage par silence **TPPS**. En fonction de l'application, une partie seulement ou la totalité des composantes du vecteur de caractéristiques $\mathbb{Z}$ est utilisée en vue d'une classification. Par exemple, si la plage de fréquences dans laquelle est extrait le spectre est compris entre 0 et 4 000 Hz, avec un pas de fréquences de 100 Hz, il est obtenu 40 éléments par vecteur spectral. Si pour la transformation du vecteur de caractéristiques brut $\mathbb{X}_i$ il est appliqué l'identité, alors sont obtenus 40 éléments pour le vecteur moyen, 40 pour le vecteur variance, et 40 pour le vecteur moment. Après concaténation et ajout des paramètres **TPPS** et **SF**, il est obtenu un vecteur de caractéristiques $\mathbb{Z}$ de 122 éléments. En fonction de l'application, il peut être choisi d'utiliser la totalité ou seulement un sous-ensemble de ce vecteur caractéristiques en prenant par exemple 40 ou 80 éléments.

Selon une variante préférée de réalisation de l'invention, le procédé consiste à assurer une opération de normalisation des composantes caractéristiques à l'aide de moyens de normalisation **45** interposés entre les moyens d'extraction **40** et le classificateur **50**. Cette normalisation consiste pour le vecteur moyen à chercher le composant qui présente la valeur maximale et à diviser les autres composants du vecteur moyen par ce maximum. Une opération similaire est effectuée pour le vecteur de variance et de moment. Pour le suivi de fréquences **SF** et le taux de passage par silence **TPPS**, ces deux paramètres sont divisés par une constante fixée après expérimentation afin d'obtenir toujours une valeur comprise entre 0,5 et 1.

Après cette étape de normalisation, il est obtenu un vecteur de caractéristiques dont chacune des composantes a une valeur comprise entre 0 et 1. Si le vecteur spectral a déjà subi une transformation, cette étape de normalisation du vecteur de caractéristiques peut ne pas être nécessaire.

Tel que cela ressort plus précisément de la **Fig. 4**, le procédé selon l'invention consiste après extraction des paramètres ou constitution des vecteurs de caractéristiques $\mathbb{Z}$, à choisir un classificateur **50** permettant à l'aide des moyens

d'identification ou de classification **60**, d'étiqueter efficacement chacun de ces vecteurs comme étant une des classes acoustiques définies.

Selon un premier exemple de réalisation, le classificateur utilisé est un réseau de neurones, tel que le perceptron multi-couches à deux couches cachées. La **Fig. 5** illustre l'architecture d'un réseau de neurones comportant par exemple 82 éléments en entrée, 39 éléments pour les couches cachées et 7 éléments en sortie. Bien entendu, il est clair que le nombre de ces éléments peut être modifié. Les éléments de la couche d'entrée correspondent aux composantes du vecteur de caractéristiques $\mathbb{Z}$. Par exemple, s'il est choisi pour la couche d'entrée 80 nœuds, il peut être utilisé une partie du vecteur de caractéristiques $\mathbb{Z}$ par exemple les composantes correspondant à la moyenne et au moment. Pour la ou les couche(s) cachée(s), les 39 éléments utilisés apparaissent suffisants, l'augmentation du nombre de neurones n'apporte pas une amélioration notable des performances. Le nombre des éléments pour la couche de sortie correspond au nombre de classes à classifier. Si deux classes sonores sont classifiées, par exemple musique et parole, la couche de sortie comporte deux nœuds.

Bien entendu, il peut être utilisé un autre type de classificateur tel que le classificateur classique K-Plus Proche Voisin (KPPV). Dans ce cas, les connaissances de l'apprentissage sont constitués simplement de données d'apprentissage. La mémorisation de l'apprentissage consiste donc à stocker toutes les données d'apprentissage. Lorsqu'un vecteur de caractéristiques $\mathbb{Z}$ se présente pour la classification, il convient de calculer les distances à toutes les données de l'apprentissage afin de choisir les classes les plus proches.

L'utilisation d'un classificateur permet d'identifier des classes sonores telles que parole ou musique, voix d'homme ou voix de femme, moment caractéristique ou moment non caractéristique d'un signal sonore, ou moment caractéristique ou moment non caractéristique accompagnant un signal vidéo au sens général représentant par exemple un film ou un match.

La description qui suit donne un exemple d'application du procédé selon l'invention pour la classification d'une bande sonore en musique ou parole. Selon cet exemple, une bande sonore en entrée est découpée en une succession d'intervalles de parole, de musique, de silence ou d'autres choses. Dans la mesure où la caractérisation d'un segment de silence est facile, les expérimentations se sont

portées sur une segmentation en parole ou en musique. Pour cette application, il a été utilisé un sous-ensemble du vecteur de caractéristiques **Z** contenant 82 éléments, 80 éléments pour la moyenne et la variance et un pour **TPPS** et un pour le **SF**. Le vecteur subit une transformation identité et une normalisation. La taille de chaque fenêtre temporelle **F** est égale à 2s.

Afin de montrer la qualité des caractéristiques ci-dessus et extraites d'un segment sonore, il a été utilisé deux classificateurs, l'un basé sur un réseau de neurone RN, l'autre utilisant le principe simple de *k*-PPV, c'est à dire « k-Plus Proche Voisin ». Dans un but de tester la généralité du procédé, il a été réalisé l'apprentissage du RN et de *k*-PPV sur 80s de musique et 80s de parole extraites de la chaîne Aljazeerah "http://www.aljazeera.net/"en langue arabe. Ensuite, les deux classificateurs ont été expérimentés sur un corpus de musique ainsi qu'un corpus de paroles, deux corpus de nature très variée totalisant 1280s (plus de 21 minutes). Le résultat sur la classification des segments de musique est donné dans le tableau suivant.

Musique extraites de	Longueur de segment	k-PPV	k-PPV % réussite	RN	RN % réussite
Apprentissage	80s	80s	100	80s	100
Fairuz (Habbaytak bissayf)	80s	74s	92.5	72s	90
Fairuz (Habbaytak bissayf)	80s	80s	100	80s	100
Fairuz (eddach kan fi nass)	80s	70s	87.5	70s	87.5
George Michael (careless whisper)	80s	70s	87.5	80s	100
George Michael (careless whisper)	80s	76s	95	80s	100
Metallica (turn the page)	80s	74s	92.5	78s	97.5
Film "Gladiateur"	80s	78s	97.5	80s	100
Total	640s	602s	94	626s	97.8

Tableau 1 taux de réussite pour la classification de musique en utilisant un RN et un *k*-PPV

On peut y voir que le classificateur k-PPV donne globalement un taux de réussite plus de 94% alors que le classificateur RN culmine avec un taux de réussite de 97,8%. On peut y noter aussi la bonne capacité de généralisation du classificateur RN. En effet, alors que l'apprentissage a été réalisé sur 80s d'une musique libanaise, il réalise une classification 100% réussie sur un genre de musique tout autre de Georges Michael et même un taux de classification réussie de 97,5% avec Metallica qui est une musique de Rock réputée difficile.

Quant à l'expérimentation sur les segments de parole, elle a été menée sur des extraits variés venant des émissions CNN en anglais, de LCI en français et du film « Gladiateur » alors que l'apprentissage des deux classificateurs a été réalisé sur 80s de parole en arabe. Le tableau suivant donne les résultats des deux classificateurs.

Paroles extraites de	Longueur de segment	k-PPV	k-PPV % réussite	RN	RN % réussite
Apprentissage	80s	80s	100	80s	100
CNN	80s	80s	100	74s	92.5
CNN	80s	72s	90	78s	97.5
CNN	80s	72s	90	76s	95
LCI	80s	58s	72.5	80s	100
LCI	80s	66s	82.5	80s	100
LCI	80s	58s	72.5	80s	100
Film "Gladiateur"	80s	72s	90	72s	90
Total	640s	558s	87.2	620s	96.9

Tableau 2 taux de réussite pour la classification de parole en utilisant un RN et un *k*-PPV

On peut voir sur le tableau que le classificateur s'avère particulièrement performant avec des extraits de LCI en français car il réalise une classification 100% correcte. Pour les extraits de CNN en anglais, il réalise tout de même un taux de bonne classification au dessus de 92,5% et globalement le classificateur RN atteint un taux de classification réussie de 97% alors que le classificateur k-PPV donne un taux de bonne classification de 87%.

Selon une autre expérience, ces résultats encourageants pour le classificateur RN a été choisi et appliqué à des segments mélangeant la parole et la musique. Pour cela, il a été réalisé un apprentissage de musique sur 40 secondes du programme « la guerre du Liban » issu de la chaîne « Aljazeerah » puis 80 secondes de parole en arabe extraites du même programme. Le classificateur RN a été testé sur 30 minutes du film "chapeau melon et bottes de cuir " qui a été segmenté et classifié. Les résultats de cette expérimentation sont donnés dans le tableau suivant.

Erreur Musique	Erreur Parole	Longueur segment	Erreur totale	Accuracy %
68s	141s	1800s	209s	88.4

Tableau 3 résultat de la segmentation-classification du film

Dans un but de comparer le classificateur selon l'invention avec les travaux de l'état de l'art, il a été aussi testé l'outil de "Muscle Fish" (http://www.musclefish.com/speechMusic.zip) utilisé par Virage sur le même corpus et les résultats suivants ont été obtenus :

Erreur Musique	Erreur Parole	Longueur segment	Erreur totale	Accuracy %
336s	36s	1800s	372s	79.3

Tableau 4 résultat de l'outil de Muscle Fish pour la segmentation-classification du film

Il peut être constaté clairement que le classificateur RN dépasse de 10 points en terme de précision l'outil Muscle Fish.

Enfin, il a été aussi testé le classificateur RN sur 10 minutes de programmes de "LCI", composés de "l'édito", de "l'invité" et de "la vie des médias" et les résultats suivants ont été obtenus :

Erreur Musique	Erreur Parole	Longueur segment	Erreur totale	Accuracy %
12s	2s	600s	14s	97.7

Tableau 5 résultat de segmentation-classification des programmes LCI

Alors que l'outil de "Muscle Fish" a donné les résultats suivants:

Erreur Musique	Erreur Parole	Longueur segment	Erreur totale	Accuracy %
2s	18s	600s	20s	96.7

Tableau 6 résultat de segmentation-classification des programmes LCI avec l'outil de Muscle Fish

Les résultats récapitulatifs par le classificateur RN sont les suivants :

Donnée d'apprentissage	Donnée de Test	Erreur totale	Apprentissage / Accuracy % test %
120s	3000s	227s	4 92.4

Tableau 7 résultat de segmentation-classification sur les différentes vidéos

On y voit que pour un taux de précision de plus de 92% sur 50 minutes dans cette expérimentation, le classificateur RN génère seulement un taux A/T (durée apprentissage/durée test) de 4 %, ce qui est très encourageant par rapport aux taux A/T de 300 % pour le système de [Will 99] (Gethin Williams, Daniel Ellis, *Speech/music discrimination based on posterior probability features*, Eurospeech 1999) basé sur les paramètres de probabilité à posteriori de HMM (Hidden Markov Model) et en utilisant les GMM.

Un deuxième exemple d'expérimentation a été réalisé afin de classifier un signal sonore en voix d'homme ou en voix de femme. Selon cette expérience, les segments de parole sont découpés en des morceaux étiquetés voix masculine ou voix féminine. A cet effet, le vecteur de caractéristiques ne comporte pas le taux de passage par silence et le suivi de fréquences. Le poids de ces deux paramètres est donc ramené à 0. La taille de la fenêtre temporelle **F** a été fixée à 1 seconde.

Les expérimentations ont été réalisées sur des données des appels téléphoniques de la base Switchboard de « Linguistic Data Consortium » LCD (http://www.ldc.upenn.edu). Il a été choisi pour l'apprentissage et pour le test des appels téléphoniques entre des locuteurs de même genre, c'est à dire conversations homme-homme et femme-femme. L'apprentissage a été fait sur 300s de parole extraites de 4 appels téléphoniques homme-homme et 300s de parole extraites de 4 appels téléphonique femme-femme. Le procédé selon l'invention a été testé sur 6000s (100min) dont 3000s extraits de 10 appels homme-homme qui sont différents des appels utilisés pour l'apprentissage, et 3000s extraits de 10 appels femme-femme, différents également des appels utilisés pour l'apprentissage. Le tableau ci-dessous résume les résultats obtenus.

Tableau 6 résultat de segmentation-classification des programmes LCI avec l'outil de Muscle Fish

Les résultats récapitulatifs par le classificateur RN sont les suivants :

Donnée d'apprentissage	Donnée de Test	Erreur totale	Apprentissage / test %	Accuracy %
120s	3000s	227s	4	92.4

Tableau 7 résultat de segmentation-classification sur les différentes vidéos

On y voit que pour un taux de précision de plus de 92% sur 50 minutes dans cette expérimentation, le classificateur RN génère seulement un taux A/T (durée apprentissage/durée test) de 4 %, ce qui est très encourageant par rapport aux taux A/T de 300 % pour le système de [Will 99] (Gethin Williams, Daniel Ellis, *Speech/music discrimination based on posterior probability features*, Eurospeech 1999) basé sur les paramètres de probabilité à posteriori de HMM (Hidden Markov Model) et en utilisant les GMM.

Un deuxième exemple d'expérimentation a été réalisé afin de classifier un signal sonore en voix d'homme ou en voix de femme. Selon cette expérience, les segments de parole sont découpés en des morceaux étiquetés voix masculine ou voix féminine. A cet effet, le vecteur de caractéristiques ne comporte pas le taux de passage par silence et le suivi de fréquences. Le poids de ces deux paramètres est donc ramené à 0. La taille de la fenêtre temporelle **F** a été fixée à 1 seconde.

Les expérimentations ont été réalisées sur des données des appels téléphoniques de la base Switchboard de « Linguistic Data Consortium » LCD (http://www.ldc.upenn.edu). Il a été choisi pour l'apprentissage et pour le test des appels téléphoniques entre des locuteurs de même genre, c'est à dire conversations homme-homme et femme-femme. L'apprentissage a été fait sur 300s de parole extraites de 4 appels téléphoniques homme-homme et 300s de parole extraites de 4 appels téléphonique femme-femme. Le procédé selon l'invention a été testé sur 6000s (100min) dont 3000s extraits de 10 appels homme-homme qui sont différents des appels utilisés pour l'apprentissage, et 3000s extraits de 10 appels femme-femme, différents également des appels utilisés pour l'apprentissage. Le tableau ci-dessous résume les résultats obtenus.

Taux de détection homme	Taux de détection femme	Longueur segment homme	Longueur segment femme	Durée de parole pour l'Apprentissage/Durée totale de test	Précision %
85%	90%	3000s	3000s	10%	87.5%

On voit que le taux de détection global est de 87,5% avec un échantillon de parole pour l'apprentissage qui n'est que de 10% des paroles testées. On constate aussi que le procédé selon l'invention réalise une meilleure détection de parole féminine (90%) que masculine (85%). Ces résultats peuvent être encore sensiblement améliorés si l'on applique le principe de vote majoritaire à des segments homogènes à la suite de la segmentation aveugle et si l'on élimine les longs silences qui apparaissent assez souvent dans les conversations téléphoniques et qui conduisent à un étiquetage de femme par la technique selon l'invention.

Une autre expérience vise à classifier un signal sonore en moment important ou non dans un match sportif. La détection de moments clés dans un match sportif par exemple celui de football dans un contexte de retransmission audiovisuel en direct est très importante pour permettre une génération automatique de résumés audiovisuels qui peuvent être une compilation des images, des moments clés ainsi détectés. Dans le contexte d'un match de football, un moment clé est celui ou intervient une action de but, une pénalité, etc. Dans le contexte d'un match de basket-ball, un moment clé peut être défini par celui où intervient une action mettant la balle dans le panier. Dans le contexte d'un match de rugby, un moment clé peut être défini par celui où intervient l'action d'essai par exemple. Cette notion de moment clé peut bien entendu être appliqué à tous matchs sportifs.

La détection de moments clés dans une séquence audiovisuelle sportive revient à un problème de la classification de la bande sonore, du terrain, de l'assistance et des commentateurs accompagnant le déroulement du match. En effet, lors des moments importants dans un match sportif, comme par exemple celui du football, ils se traduisent en une tension dans le ton de parole du commentateur et l'intensification du bruit des spectateurs. Devant cette expérimentation, le vecteur de

Taux de détection homme	Taux de détection femme	Longueur segment homme	Longueur segment femme	Durée de parole pour l'Apprentissage / Durée totale de test	Précision %
85%	90%	3000s	3000s	10%	87.5%

On voit que le taux de détection global est de 87,5% avec un échantillon de parole pour l'apprentissage qui n'est que de 10% des paroles testées. On constate aussi que le procédé selon l'invention réalise une meilleure détection de parole féminine (90%) que masculine (85%). Ces résultats peuvent être encore sensiblement améliorés si l'on applique le principe de vote majoritaire à des segments homogènes à la suite de la segmentation aveugle et si l'on élimine les longs silences qui apparaissent assez souvent dans les conversations téléphoniques et qui conduisent à un étiquetage de femme par la technique selon l'invention.

Une autre expérience vise à classifier un signal sonore en moment important ou non dans un match sportif. La détection de moments clés dans un match sportif par exemple celui de football dans un contexte de retransmission audiovisuel en direct est très importante pour permettre une génération automatique de résumés audiovisuels qui peuvent être une compilation des images, des moments clés ainsi détectés. Dans le contexte d'un match de football, un moment clé est celui ou intervient une action de but, une pénalité, etc. Dans le contexte d'un match de basket-ball, un moment clé peut être défini par celui où intervient une action mettant la balle dans le panier. Dans le contexte d'un match de rugby, un moment clé peut être défini par celui où intervient l'action d'essai par exemple. Cette notion de moment clé peut bien entendu être appliqué à tous matchs sportifs.

La détection de moments clés dans une séquence audiovisuelle sportive revient à un problème de la classification de la bande sonore, du terrain, de l'assistance et des commentateurs accompagnant le déroulement du match. En effet, lors des moments importants dans un match sportif, comme par exemple celui du football, ils se traduisent en une tension dans le ton de parole du commentateur et l'intensification du bruit des spectateurs. Devant cette expérimentation, le vecteur de

caractéristiques utilisé est celui utilisé pour la classification musique/parole en enlevant uniquement les deux paramètres **TPPS** et de **SF**. La transformation utilisée sur les vecteurs de caractéristiques bruts est celle suivant l'échelle de Mel, tandis que l'étape de la normalisation n'est pas appliquée au vecteur de caractéristiques. La taille de la fenêtre temporelle **F** est de 2 secondes.

Il a été choisi trois matchs de football de la coupe de l'UEFA pour les expérimentations. Pour l'apprentissage, il a été segmenté manuellement 20s des moments clés, et 20s des moments non clés du premier match. On a donc deux classes sonores : moment clé ou moment non clé.

Après l'apprentissage, il a été mené la classification sur les trois matchs. Les résultats sont évalués en terme du nombre de buts détectés, et en terme du temps classifié comme important.

	Nombre de buts	Temps important détecté (s)	Buts détectés	Précision %
Match 1	3	90	3	100
Match 2	0	40	0	NA
Match 3	4	80	4	100

On peut voir qu'à travers le tableau, tous les moments de but ont été détectés. En plus, pour un match de football de 90 minutes, on génère un résumé de 90 secondes au plus comprenant tous les moments de but.

Bien entendu, la classification en moments importants ou non peut être généralisée à la classification sonore de tous documents audiovisuels, tels qu'un film d'action ou un film pornographique.

Le procédé selon l'invention permet également par tous moyens appropriés, d'affecter une étiquette pour chaque fenêtre temporelle affectée à une classe et de rechercher les étiquettes pour un tel signal sonore par exemple enregistré dans une base de données.

L'invention n'est pas limitée aux exemples décrits et représentés car diverses modifications peuvent y être apportées sans sortir de son cadre.

caractéristiques utilisé est celui utilisé pour la classification musique/parole en enlevant uniquement les deux paramètres **TPPS** et de **SF**. La transformation utilisée sur les vecteurs de caractéristiques bruts est celle suivant l'échelle de Mel, tandis que l'étape de la normalisation n'est pas appliquée au vecteur de caractéristiques. La taille de la fenêtre temporelle **F** est de 2 secondes.

Il a été choisi trois matchs de football de la coupe de l'UEFA pour les expérimentations. Pour l'apprentissage, il a été segmenté manuellement 20s des moments clés, et 20s des moments non clés du premier match. On a donc deux classes sonores : moment clé ou moment non clé.

Après l'apprentissage, il a été mené la classification sur les trois matchs. Les résultats sont évalués en terme du nombre de buts détectés, et en terme du temps classifié comme important.

	Nombre de buts	Temps important détecté (s)	Buts détectés	Précision %
Match 1	3	90	3	100
Match 2	0	40	0	NA
Match 3	4	80	4	100

On peut voir qu'à travers le tableau, tous les moments de but ont été détectés. En plus, pour un match de football de 90 minutes, on génère un résumé de 90 secondes au plus comprenant tous les moments de but.

Bien entendu, la classification en moments importants ou non peut être généralisée à la classification sonore de tous documents audiovisuels, tels qu'un film d'action ou un film pornographique.

Le procédé selon l'invention permet également par tous moyens appropriés, d'affecter une étiquette pour chaque fenêtre temporelle affectée à une classe et de rechercher les étiquettes pour un tel signal sonore par exemple enregistré dans une base de données.

L'invention n'est pas limitée aux exemples décrits et représentés car diverses modifications peuvent y être apportées sans sortir de son cadre.

REVENDICATIONS

1 - Procédé pour affecter au moins une classe sonore à un signal sonore, caractérisé en ce qu'il comprend les étapes suivantes :

- diviser le signal sonore en des segments temporels (T) présentant une durée déterminée,
- extraire les paramètres fréquentiels du signal sonore dans chacun des segments temporels (T),
- regrouper les paramètres fréquentiels dans des fenêtres temporelles (F) présentant une durée déterminée supérieure à la durée des segments temporels (T),
- extraire de chaque fenêtre temporelle (F), des composantes caractéristiques,
- et en considération des composantes caractéristiques extraites et à l'aide d'un classificateur, identifier la classe sonore des fenêtres temporelles (F) du signal sonore.

2 - Procédé selon la revendication 1, caractérisé en ce qu'il consiste à diviser le signal sonore en des segments temporels (T) dont la durée est comprise entre 10 et 30 ms.

3 - Procédé selon la revendication 1, caractérisé en ce qu'il consiste à extraire les paramètres fréquentiels du signal sonore en déterminant une série des valeurs du spectre de fréquence dans une plage de fréquences comprise entre une fréquence minimale et une fréquence maximale.

4 - Procédé selon la revendication 3, caractérisé en ce qu'il consiste à extraire les paramètres fréquentiels en utilisant la Transformée de Fourier Discrète.

5 - Procédé selon la revendication 3 ou 4, caractérisé en ce qu'il consiste à assurer une opération de transformation ou de filtrage des paramètres fréquentiels.

6 - Procédé selon la revendication 5, caractérisé en ce qu'il consiste à réaliser une transformation de type identité, moyenne de deux fréquences adjacentes, ou selon l'échelle de Mel.

7 - Procédé selon l'une des revendications 3 à 5, caractérisé en ce qu'il consiste à regrouper les paramètres fréquentiels dans des fenêtres temporelles de durée supérieure à 0,3 seconde et de préférence comprise entre 0,5 et 2 secondes.

8 - Procédé selon la revendication 1, caractérisé en ce qu'il consiste à extraire de chaque fenêtre temporelle, des composantes caractéristiques telles que la moyenne, la variance, le moment, le paramètre du suivi des fréquences ou le taux de passage par silence.

9 - Procédé selon la revendication 8, caractérisé en ce qu'il consiste à utiliser une ou plusieurs composantes caractéristiques en entrée du classificateur.

10 - Procédé selon la revendication 8 ou 9, caractérisé en ce qu'il consiste à assurer une opération de normalisation des composantes caractéristiques.

11 - Procédé selon les revendications 8 et 10, caractérisé en ce que l'opération de normalisation consiste :

- pour la moyenne, la variance ou le moment, chercher le composant présentant la valeur maximale et à diviser les autres composants par cette valeur maximale,
- pour le suivi des fréquences ou le taux de passage par silence, à diviser chacune de ces composantes caractéristiques par une constante fixée après expérimentation pour obtenir une valeur comprise entre 0,5 et 1.

12 - Procédé selon la revendication 1 ou 9, caractérisé en ce qu'il consiste à utiliser comme classificateur, un réseau de neurones ou le K-Plus Proche Voisin.

13 - Procédé selon la revendication 12, caractérisé en ce qu'il consiste à réaliser une phase d'apprentissage d'un signal sonore pour le classificateur.

14 - Procédé selon les revendications 1 à 13, caractérisé en ce qu'il consiste à l'aide d'un classificateur, à identifier des classes sonores telles que parole ou musique, voix d'homme ou voix de femme, moment caractéristique ou moment non caractéristique d'un signal sonore, moment caractéristique ou moment non caractéristique accompagnant un signal vidéo représentant, par exemple, un film ou un match.

15 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à classifier le signal sonore en musique ou en parole en utilisant les paramètres de moyenne, de variance, de suivi de fréquences, et le taux de passage par silence, suivi par une normalisation des paramètres tandis que la fenêtre temporelle est égale à 2 s.

16 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à classifier le signal d'un match en moment important ou moment non important en utilisant les

paramètres de moyenne et de variance, avec une transformation selon l'échelle de Mel sans appliquer une normalisation des composantes caractéristiques.

17 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à identifier des moments forts dans un signal sonore d'un match.

18 - Procédé selon la revendication 17, caractérisé en ce qu'il consiste à utiliser l'identification des moments forts pour créer un résumé de match.

19 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à identifier et suivre la parole dans un signal sonore.

20 - Procédé selon la revendication 19, caractérisé en ce qu'il consiste à identifier et suivre la parole d'un homme et/ou d'une femme pour la partie parole du signal sonore.

21 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à identifier et suivre la musique dans un signal sonore.

22 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à déterminer si le signal sonore contient de la parole ou de la musique.

23 - Procédé selon la revendication 14, caractérisé en ce qu'il consiste à affecter une étiquette pour chaque fenêtre temporelle affectée à une classe.

24 - Procédé selon la revendication 23, caractérisé en ce qu'il consiste à rechercher les étiquettes pour un signal sonore.

25 - Appareil pour affecter au moins une classe sonore à un signal sonore, caractérisé en ce qu'il comprend :

- des moyens (**10**) pour diviser le signal sonore (S) en des segments temporels (**T**) présentant une durée déterminée,
- des moyens (**20**) pour extraire les paramètres fréquentiels du signal sonore dans chacun des segments temporels (**T**),
- des moyens (**30**) pour regrouper les paramètres fréquentiels dans des fenêtres temporelles (**F**) présentant une durée déterminée supérieure à la durée des segments temporels,
- des moyens (**40**) pour extraire de chaque fenêtre temporelle (**F**), des composantes caractéristiques,

- et des moyens (**60**) pour identifier la classe sonore des fenêtres temporelles (**F**) du signal sonore en considération des composantes caractéristiques extraites et à l'aide d'un classificateur.

26 - Appareil selon la revendication 25, caractérisé en ce que les moyens (**20**) pour extraire les paramètres fréquentiels utilisent la Transformée de Fourier Discrète.

27 - Appareil selon la revendication 25 ou 26, caractérisé en ce qu'il comprend des moyens (**25**) pour assurer une opération de transformation ou de filtrage des paramètres fréquentiels.

28 - Appareil selon l'une des revendications 24 à 27, caractérisé en ce qu'il comporte des moyens (**30**) pour regrouper les paramètres fréquentiels dans des fenêtres temporelles (**F**) de durée supérieure à 0,3 seconde et de préférence comprise entre 0,5 et 2 secondes.

29 - Appareil selon la revendication 2, caractérisé en ce qu'il comporte en tant que moyens (**40**) pour extraire de chaque fenêtre temporelle, des composantes caractéristiques, des moyens pour extraire la moyenne, la variance, le moment, le paramètre du suivi des fréquences ou le taux de passage par silence.

30 - Appareil selon la revendication 29, caractérisé en ce qu'il comporte des moyens (**45**) de normalisation des composantes caractéristiques.

31 - Appareil selon la revendication 24, caractérisé en ce qu'il comporte comme classificateur, un réseau de neurones ou le K-Plus Proche Voisin.

32 - Appareil selon la revendication 1, caractérisé en ce qu'il comprend des moyens (**60**) pour identifier des classes sonores telles que parole ou musique, voix d'homme ou voix de femme, moment caractéristique ou moment non caractéristique d'un signal sonore, moment caractéristique ou moment non caractéristique accompagnant un signal vidéo représentant, par exemple, un film ou un match.

33 - Appareil selon la revendication 24, caractérisé en ce qu'il comporte des moyens pour affecter une étiquette pour chaque fenêtre temporelle affectée à une classe.

34 - Appareil selon la revendication 33, caractérisé en ce qu'il comprend des moyens pour rechercher les étiquettes pour un signal sonore enregistré dans une base de données.

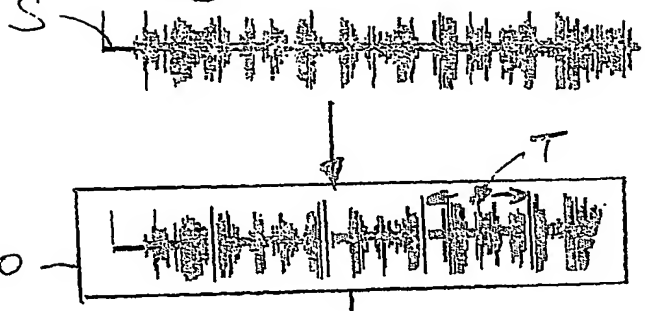
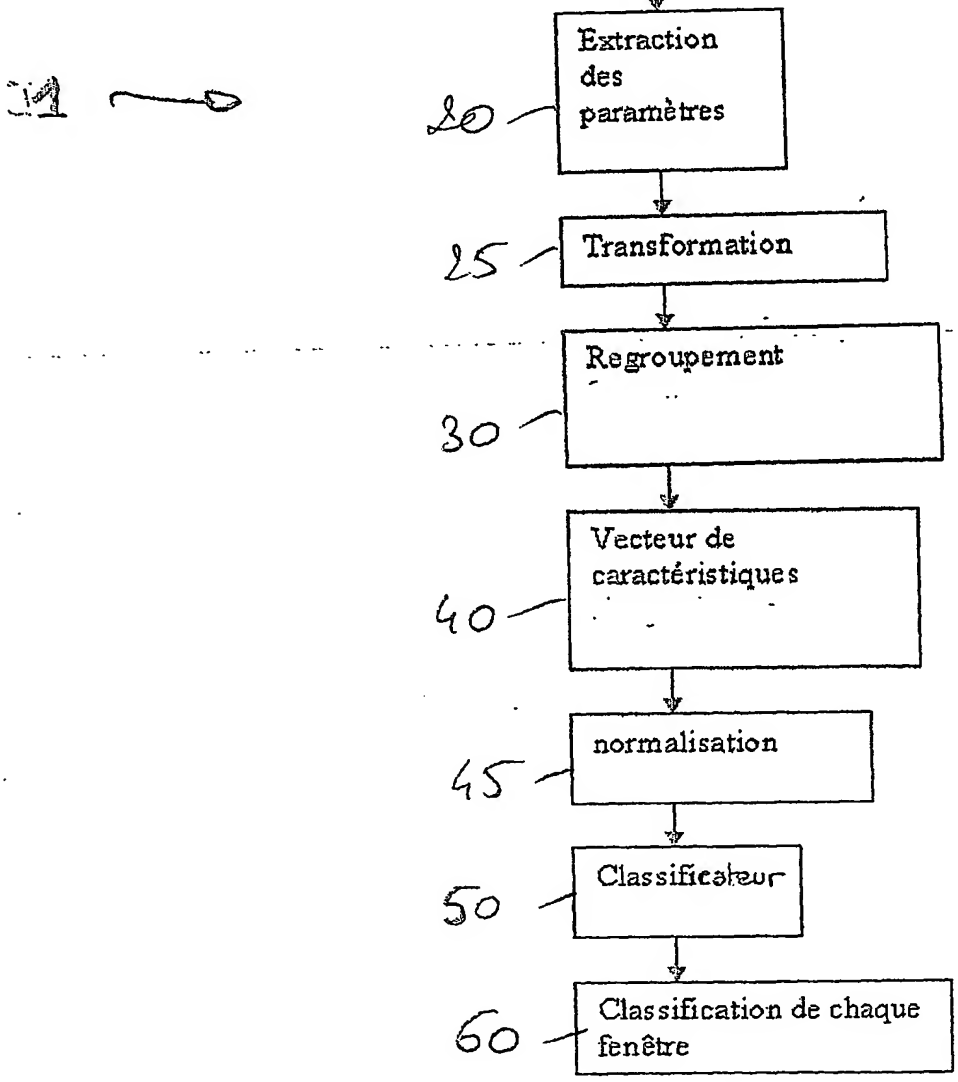

S
T
10

FIG 1


Extraction des paramètres
20
Transformation
25
Regroupement
30
Vecteur de caractéristiques
40
normalisation
45
Classificateur
50
Classification de chaque fenêtre
60

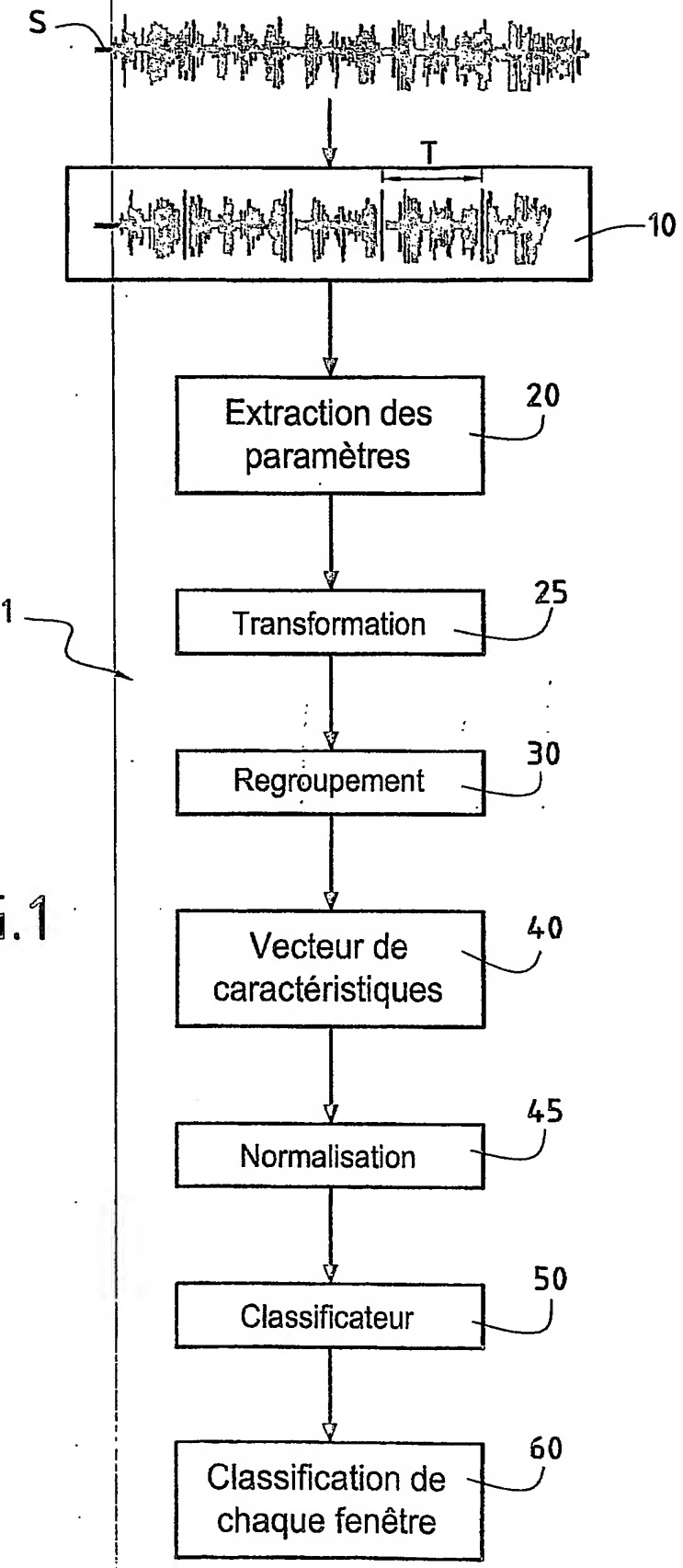
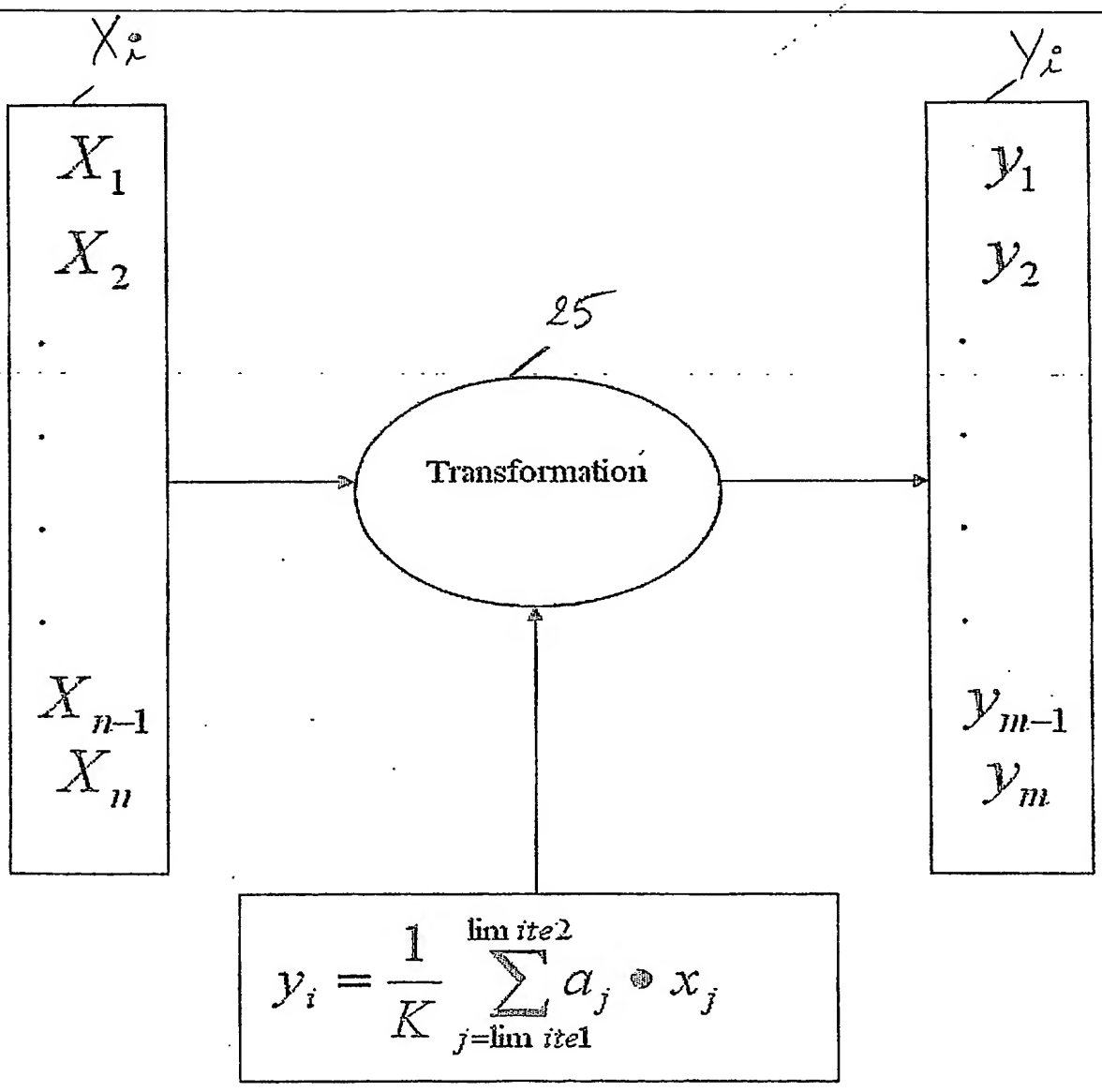

S
T
10
Extraction des paramètres
20
1
Transformation
25
Regroupement
30
Vecteur de caractéristiques
40
Normalisation
45
Classificateur
50
Classification de chaque fenêtre
60

FIG.1

FIG 2


Xi
X1
X2
Xn-1
Xn
25
Transformation
Yi
y1
y2
ym-1
ym
$y_i = \frac{1}{K} \sum_{j=\lim ite1}^{\lim ite2} a_j \bullet x_j$

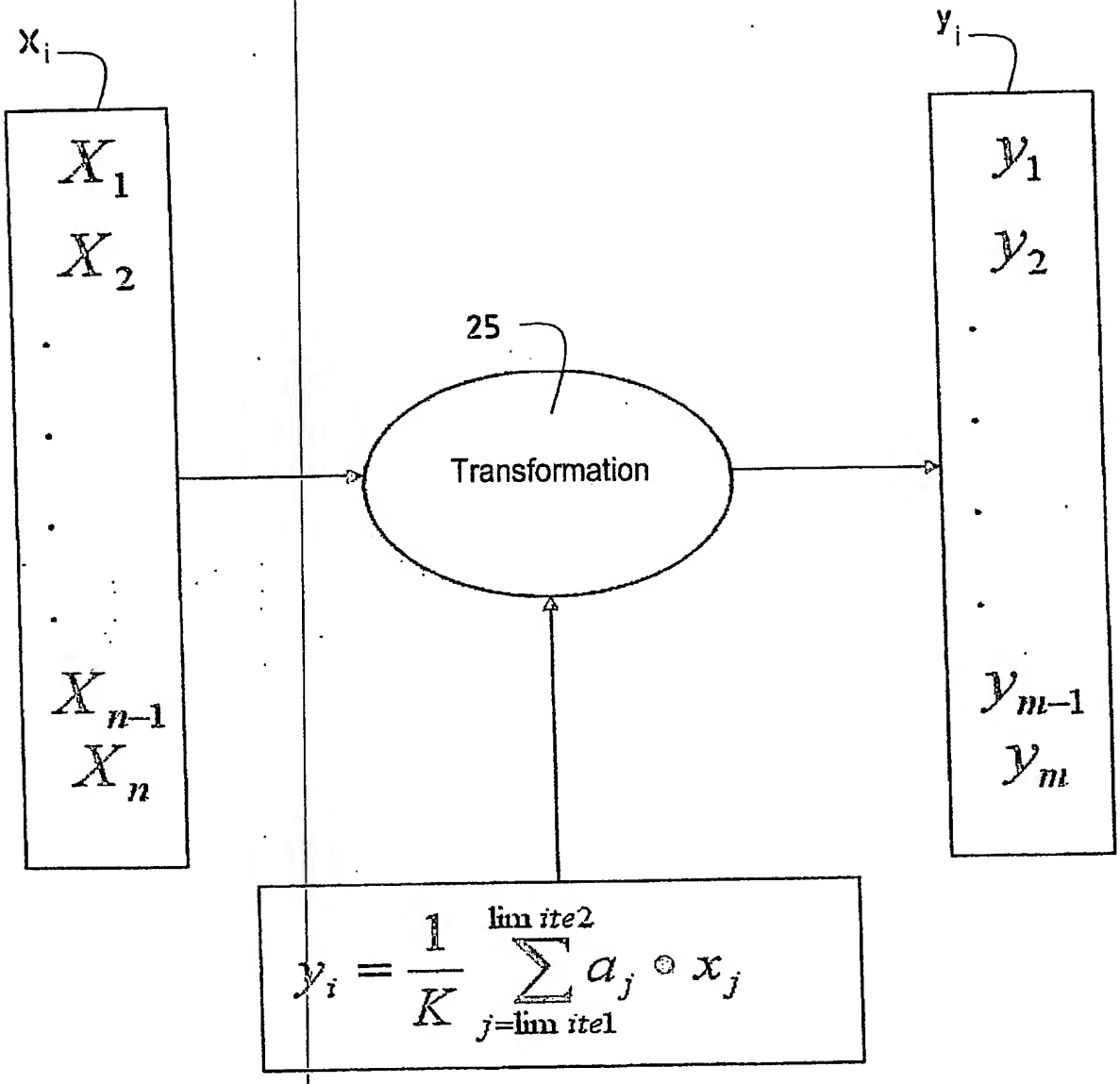
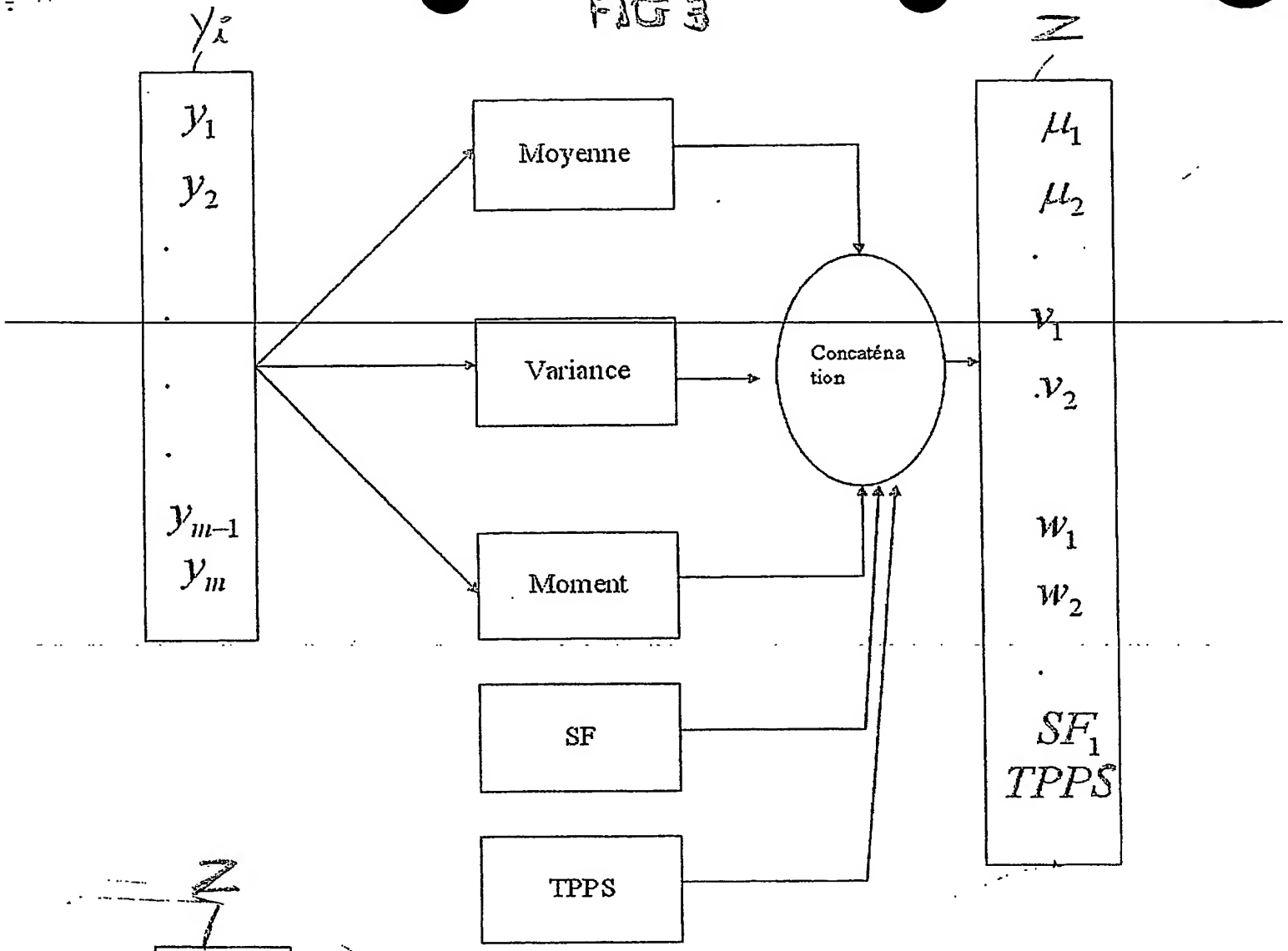

x_i
X_1
X_2
X_{n-1}
X_n
25
Transformation
y_i
y_1
y_2
y_{m-1}
y_m
$y_i = \frac{1}{K} \sum_{j=\lim ite1}^{\lim ite2} a_j \circ x_j$

FIG.2

FIG 3


y_i
y_1
y_2
y_{m-1}
y_m
Moyenne
Variance
Moment
SF
TPPS
Concaténation
Z
μ_1
μ_2
v_1
v_2
w_1
w_2
SF_1
TPPS

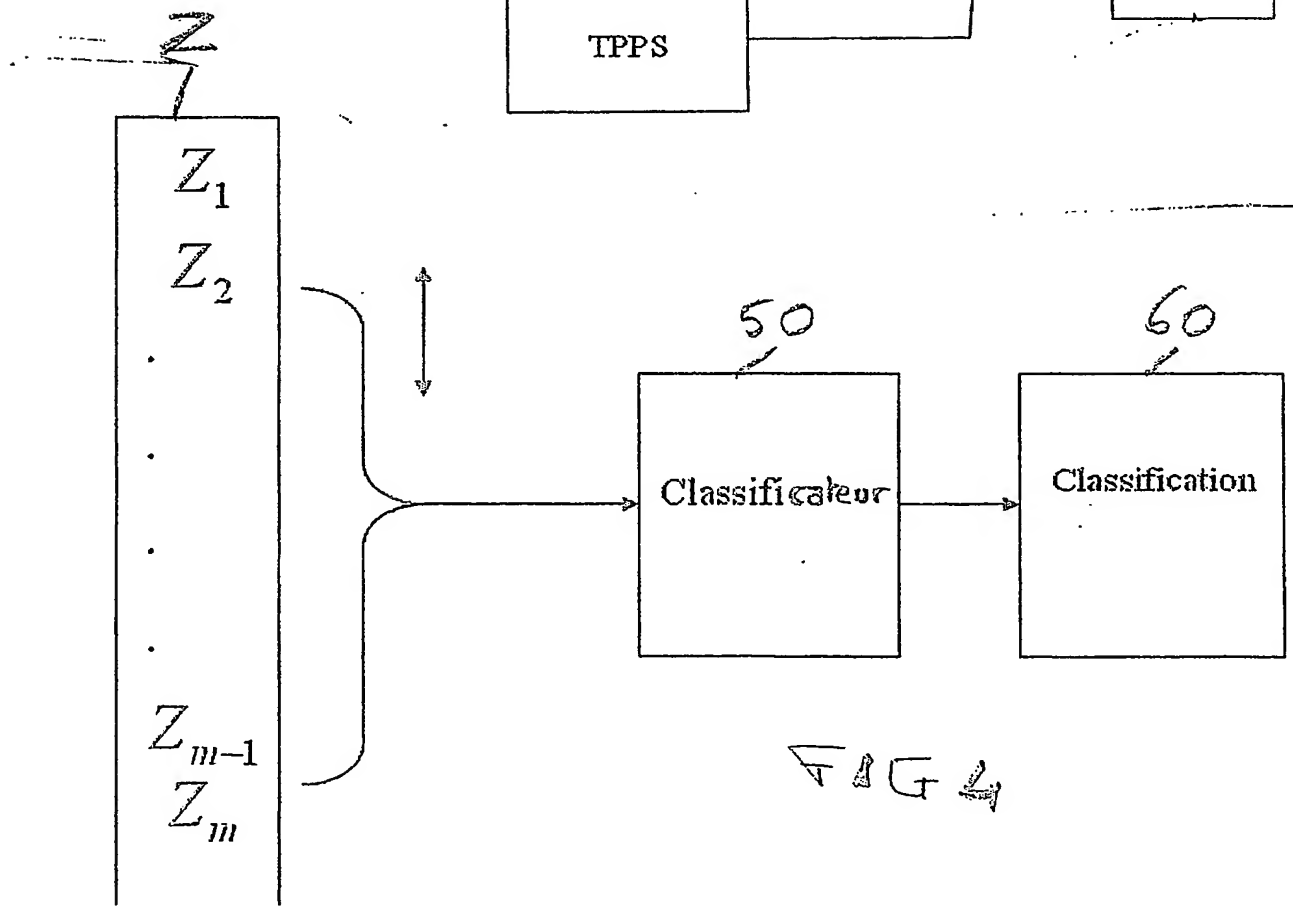

Z
Z_1
Z_2
Z_{m-1}
Z_m
50
Classificateur
60
Classification

FIG 4

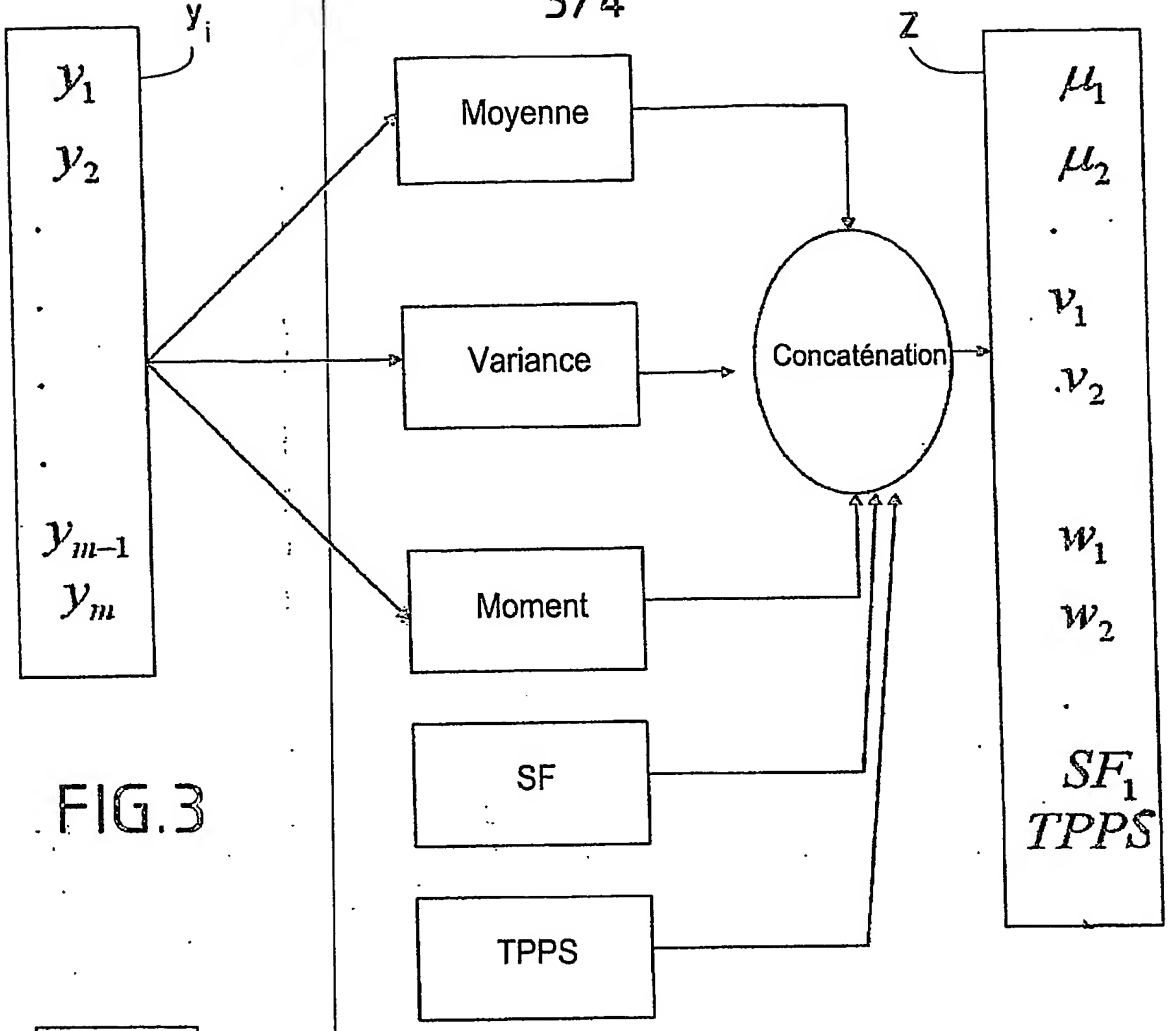

y_i
y_1
y_2
y_{m-1}
y_m
Moyenne
Variance
Moment
SF
TPPS
Concaténation
z
μ_1
μ_2
v_1
v_2
w_1
w_2
SF_1
$TPPS$

FIG.3

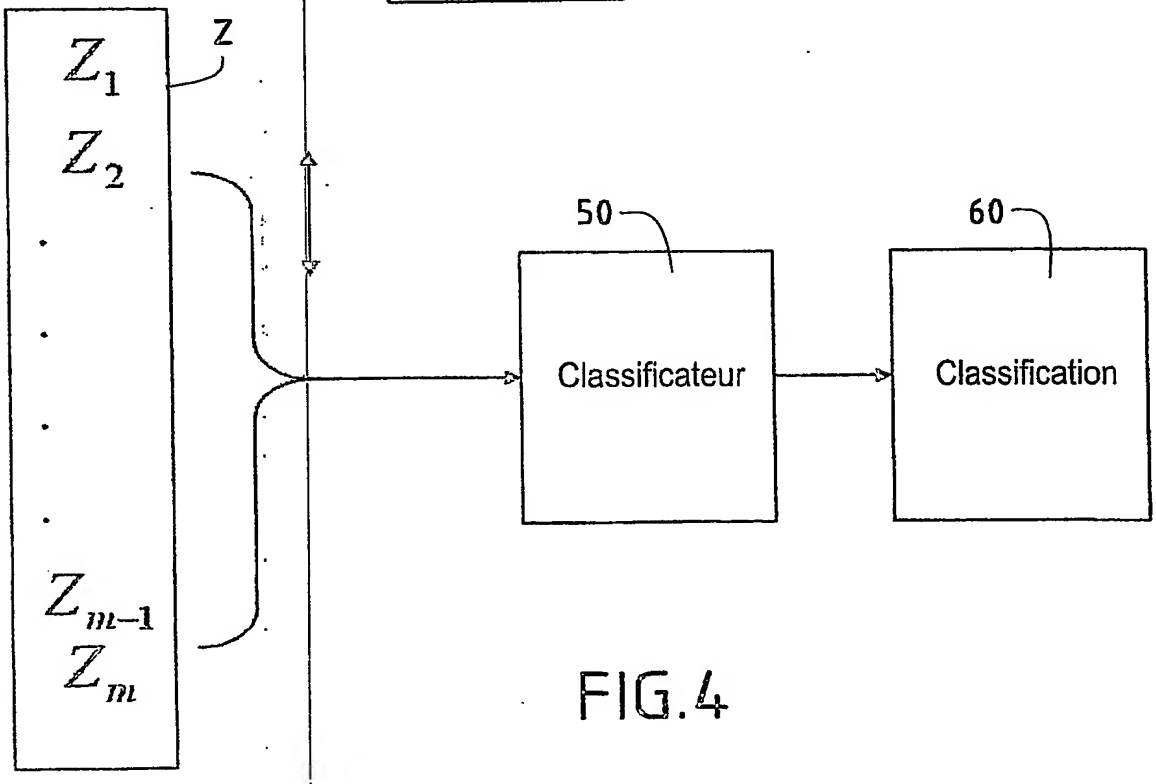
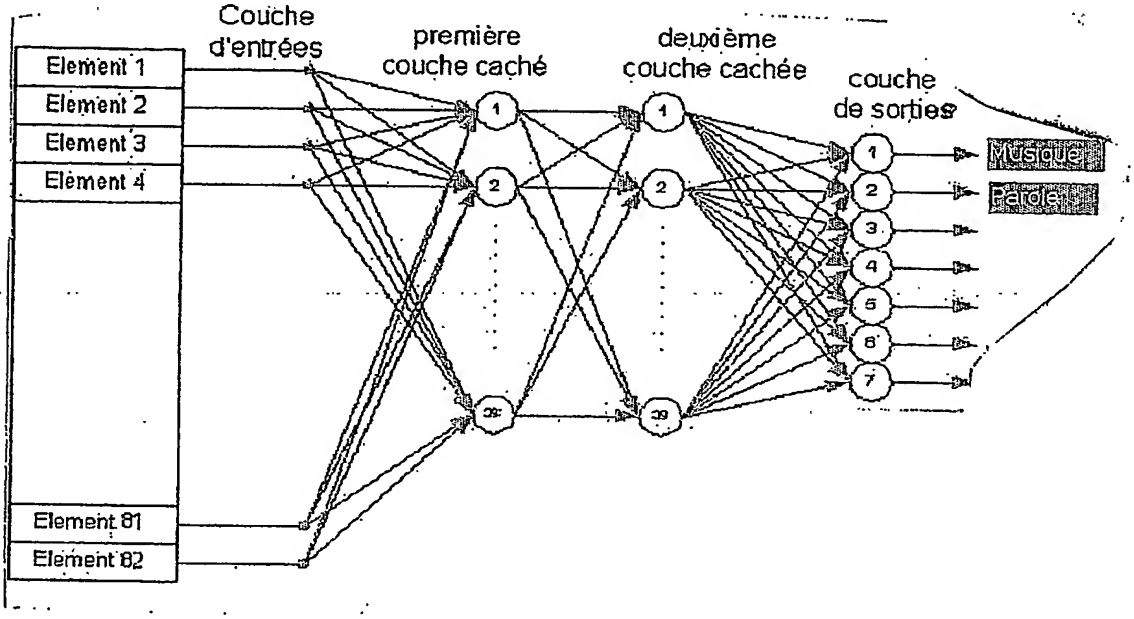

z
Z_1
Z_2
Z_{m-1}
Z_m
50
Classificateur
60
Classification

FIG.4

FIG 5


Couche d'entrées
première couche caché
deuxième couche cachée
couche de sorties
Element 1
Element 2
Element 3
Element 4
Element 81
Element 82
1
2
39
1
2
39
1
2
3
4
5
6
7
Musique
Parole

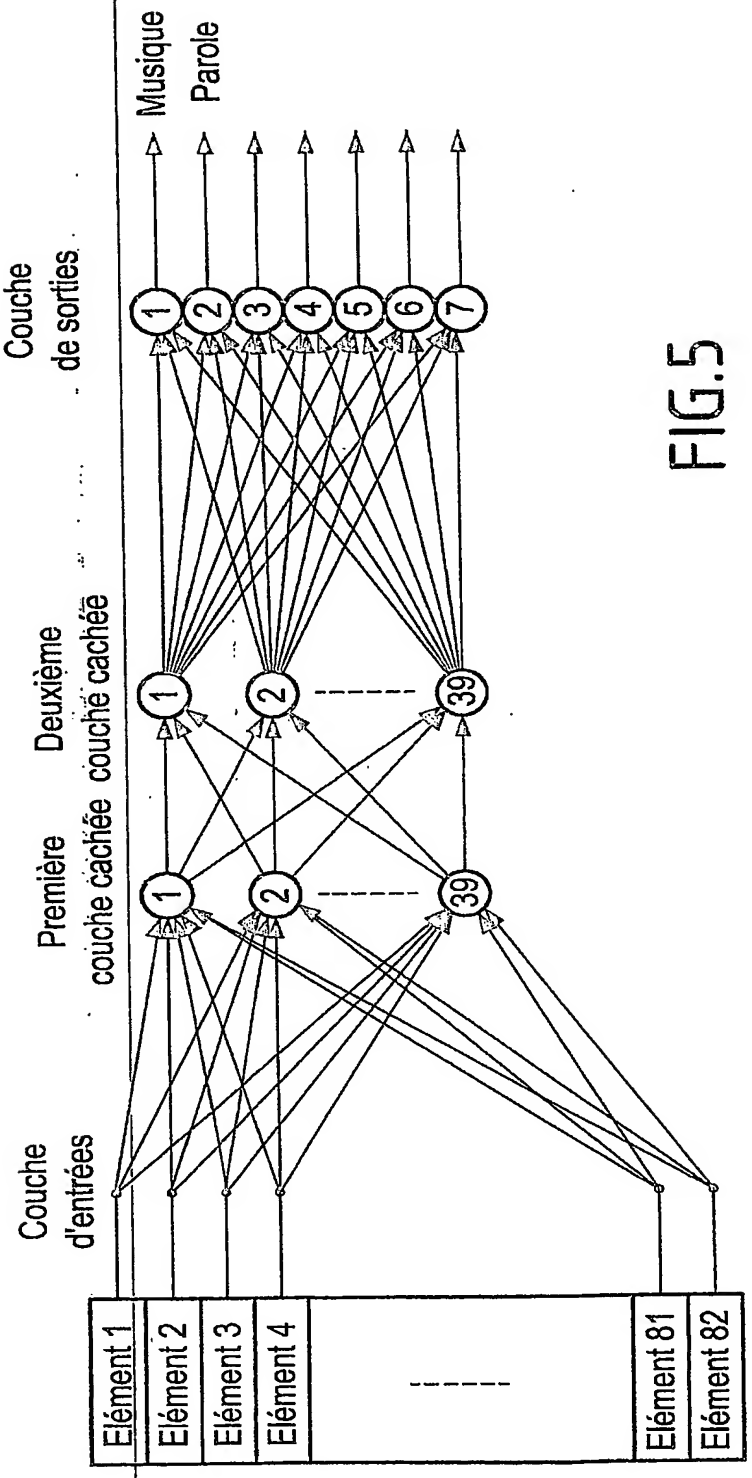

Musique
Parole
Couche de sorties
Deuxième couche cachée
Première couche cachée
Couche d'entrées
Elément 1
Elément 2
Elément 3
Elément 4
Elément 81
Elément 82

FIG.5

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11235*02

DÉPARTEMENT DES BREVETS
26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 93 59 30

DÉSIGNATION D'INVENTEUR(S) Page N° 1 / 1

(Si le demandeur n'est pas l'inventeur ou l'unique inventeur)

Cet imprimé est à remplir lisiblement à l'encre noire

DB 113 W /260899

Vos références pour ce dossier *(facultatif)*	70416BFR23 JMT/VF
N° D'ENREGISTREMENT NATIONAL	0208518

TITRE DE L'INVENTION (200 caractères ou espaces maximum)

PROCEDE ET APPAREIL POUR AFFECTER UNE CLASSE SONORE A UN SIGNAL SONORE

LE(S) DEMANDEUR(S) :

Cabinet BEAU DE LOMENIE
51, Avenue Jean Jaurès
B.P. 7073
69301 LYON CEDEX 07

DESIGNE(NT) EN TANT QU'INVENTEUR(S) : (Indiquez en haut à droite «Page N° 1/1» S'il y a plus de trois inventeurs, utilisez un formulaire identique et numérotez chaque page en indiquant le nombre total de pages).

Nom		HARB	
Prénoms		Hadi	
Adresse	Rue	19, Rue de la Victoire	
	Code postal et ville	69003	LYON
Société d'appartenance *(facultatif)*			
Nom		CHEN	
Prénoms		Liming	
Adresse	Rue	45, Boulevard des Brotteaux	
	Code postal et ville	69006	LYON
Société d'appartenance *(facultatif)*			
Nom			
Prénoms			
Adresse	Rue		
	Code postal et ville		
Société d'appartenance *(facultatif)*			

DATE ET SIGNATURE(S)
DU (DES) DEMANDEUR(S)
OU DU MANDATAIRE
(Nom et qualité du signataire)

LYON, le 8 Juillet 2002
J-M. THIBAULT
Conseil en P.I. - N° 94-0312

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire. Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.